Autopiquer - a Robust and Reliable Peak Detection Algorithm for Mass Spectrometry

Autopiquer - a robust and reliable peak detection algorithm for mass spectrometry

Running title: Autopiquer peak detection

Address reprint requests to: David P A Kilgour, Department of Chemistry and Forensics, Nottingham Trent University, Nottingham, NG11 8NS, UK; +44(0)115 84 83403; david.kilgour@ntu.ac.uk

Autopiquer - a robust and reliable peak detection algorithm for mass spectrometry

David P.A. Kilgour[*1], *Sam Hughes*[2], *Samantha L. Kilgour*[3], *C. Logan Mackay*[2], *Magnus Palmblad*[4], *Bao Quoc Tran*[5], *Young Ah Goo*[5], *Robert K. Ernst*[3], *David J. Clarke*[2], *and David R. Goodlett*[5]

[1] Department of Chemistry and Forensics, Nottingham Trent University, Nottingham, NG11 8NS, UK.

[2] EastCHEM School of Chemistry, University of Edinburgh, Edinburgh, EH9 3FJ, UK.

[3] Department of Microbial Pathogenesis, School of Dentistry, University of Maryland Baltimore, 21201,Maryland, USA.

[4] Center for Proteomics and Metabolomics, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands.

[5] Department of Pharmaceutical Sciences, School of Pharmacy, University of Maryland Baltimore, 21201, Maryland, USA.

[*] Corresponding author. Email: david.kilgour@ntu.ac.uk

ABSTRACT: We present a simple algorithm for robust and unsupervised peak detection by determining a noise threshold in isotopically resolved mass spectrometry data. Solving this problem will greatly reduce the subjective and time consuming manual picking of mass spectral peaks and so will prove beneficial in many research applications. The Autopiquer approach uses autocorrelation to test for the presence of (isotopic) structure in overlapping windows across the spectrum. Within each window, a noise threshold is optimized to remove the most unstructured data whilst keeping as much of the (isotopic) structure as possible. This algorithm has been successfully demonstrated for both peak detection and spectral compression on data from many different classes of mass spectrometer and for different sample types and this approach should also be extendible to other types of data that contain regularly spaced discrete peaks.

INTRODUCTION

Peak detection is a key stage in the extraction of information from mass spectral data. Despite the fundamental importance of this data processing step, it is not yet a solved problem. There are many different existing methods for producing a peak detection threshold in mass spectrometry. Some of the most widely known are (i) the spectral mean + $n \times$ spectral or noise standard deviation (sometimes called n-Sigma); [1-3] (ii) the mean of the local maxima; [4] (iii) the valley between bi- or multi-modal spectral intensity distributions;[5] (iv) the signal-to-noise calculation using data between isotopic peaks;[6] and (v) the simple threshold derived from the root mean square of the spectrum.

In practice, use of any of these methods for peak detection purposes in mass spectra requires a level of expertise on the behalf of the user and will often, even in skilled hands, produce results that can cause many peaks in a spectrum to be missed. This weakness is a considerable disadvantage when dealing with spectra which may contain important but low signal-to-noise ratio (SNR) peaks that may be interspersed amongst much more intense peaks. For this reason, it is a commonplace occurrence for mass spectrometry users to resort to manual peak detection methods – a process that is very time consuming and will inevitably result in an unwanted introduction of a degree of user subjectivity/accidental bias that will reduce the ability to reliably compare data.

Furthermore, as mass spectrometry datasets become ever larger (for example from mass spectrometry imaging and large population metabolomics applications) it becomes increasingly impractical to manually peak pick these data, to achieve the same quality of results as can be achieved when working only with single spectra.

We wished to develop a method for robustly and reliably detecting peaks in mass spectral data that would be simple to use, would require minimal user input and would outperform current standard methods for peak detection providing results closer in coverage to manual peak detection. We also wanted to develop a method that would be sufficiently fast to work within existing workflows.

Within spectra, those components of the data that can be used to derive useful information are termed the signal. Conversely, those components that do not contain useful information are noise. In order for the signal to be detectable, it must exhibit an intensity such that it is visible above the noise. Developing an understanding of a threshold level that will be used to provide an indication of whether a point is considered to be more likely to result from signal or more likely to result from noise is fundamental to being able to extract information from a mass spectrum in two ways: firstly because this threshold will be used to detect peak regions (as those regions will be above the threshold) and secondly because the relative height of the peak will be expressed in terms of the signal to noise ratio (SNR).

The noise threshold in mass spectra is often not constant across the spectrum. Therefore, it is common practise to determine the threshold in small sections of the spectrum (termed windows). These windows can be discrete (a series of adjacent windows of a certain width; the threshold value calculated for that window is applied to all points within it) or moving (a window of a certain width is calculated for and centred on every point in turn, and the threshold produced from each window is used for one point only).

Estimating a robust level that can be used to distinguish signal and noise is not trivial. The recognition of peaks in a noisy spectrum is a task, similar to the detection of faces, that humans find simple but that is difficult for computers.[7] If the peaks in a spectrum are visible at greater intensity than the noise, then the signal and the noise must have different statistical distributions. Therefore, some approaches to peak detection have attempted to define the statistical distribution of the noise within a spectrum in order to be able to identify the peaks as anomalies relative to that noise distribution. The first difficulty with this approach comes from the fact that the system does not know (ab initio) if the spectrum under consideration contains only a few peaks or if it is very peak dense. Therefore, it is difficult to be able to automatically select a spectral region (within any window – remembering that the noise distribution is often not constant across a spectrum) that contains no peaks or at least is mostly noise, in order to be able to derive some statistical information of the background noise that could be used to estimate the likelihood of a point of a given intensity being part of the noise distribution within that window. Thus, the statistics are calculated from the entire spectrum in a window. If the statistics for the noise are calculated from a window that contains a significant proportion that is actually part of a peak or peaks, then the estimation of the upper end of the noise distribution will be over-estimated and the threshold will be set too high, leading to missed peaks. This is a common problem found using the *n*-Sigma method[2,3] of estimating the noise threshold as well as the mean of the local maxima method[4] and the signal-to-noise calculation using data between isotopic peaks, as described in Horn et al.[6]

The second difficulty arising from unsupervised whole spectrum statistical approaches is that in most mass spectrometry data, the signal distribution and the noise distribution overlap to such a degree that the two cannot be completely separated by intensity alone. This feature further

complicates any efforts to use statistics associated with the spectral intensities alone to separate the signal from the noise. It must be noted that for such spectra where there is a partial separation of the signal and noise distributions as a function of intensity, there is an easy to implement method for identifying a peak threshold [5] – however spectra of this type appear to be rare.

Given the difficulties associated with the statistical methods based on intensity alone, we looked for another approach. Conveniently, in mass spectrometry data there is another feature, besides greater intensity, that will be exhibited by real peaks: as a consequence of the natural abundances of the various stable isotopes of the common elements, mass spectral data will contain a series of isotopologue peaks for most ions – commonly known as an isotopic distribution. Falsely detected noise peaks will not (or are at least very unlikely to) exhibit this predictable structure. By seeking to isolate this pattern, one can develop a method to identify an optimum threshold between the noise and signal components in a spectrum.

There have been previously presented methods for peak picking in mass spectral data that have included information about the isotopic distribution in the methodology. A well described example of this is the Isotope Wavelet method, developed by Hussong et al.[8,9], or the Sophisticated Numerical Annotation Procedure (SNAP) algorithm used by Bruker.[10] These methods rely on there being a scalable isotopic distribution that can be applied to all peaks. In the case of the isotopic wavelet, the isotopic distribution is estimated for peptides using the averagine model.[11] This approach will work well for standard peptides but often fails to provide robust peak detection in many applications, when there are less well defined or no generalizable isotopic models that can be used; for example, metalloproteins (or other metal-ligand complexes),

dissolved organic matter, isotopically enriched or depleted samples etc. Most importantly, as with many peak detection techniques described in the literature, the isotope wavelet and other wavelet methods often rely on the user supplying the peak detection threshold – and our aim was to avoid this requirement.[9,12,13] However, it would be possible to use the threshold we propose below for this step, in these isotopic pattern based peak picking algorithms.

We have developed an alternate method for peak detection, known as Autopiquer, that is based on the expectation that real peaks should display regular (isotopic) spacing in a mass spectrum whereas contributions from noise will not. The new method uses autocorrelation to detect regular patterns within equally spaced windows across the spectrum. While autocorrelation has been used previously as a method for providing increased confidence on detected peaks, for example by Palmblad et al.[14], it has not, so far as we know, been used as a means of proposing a peak detection threshold.

As a consequence of the subjective user input required in most peak detection methods it is impossible to provide objective metrics to describe the performance of such techniques. Furthermore, with the wide variety of spectral filtering, smoothing and peak picking methodologies out there, it would be very difficult to provide metrics to cover all possible combinations of procedure. Therefore, we will concentrate on providing illustrations of the performance of the Autopiquer algorithm when used on raw data. Here, we analyse complex mass spectral datasets obtained during top-down protein fragmentation (fragmentation of a single intact protein).[15] These spectra are known to consist of many hundreds of fragment ions (that occur over large mass, charge and intensity ranges) with many overlapping isotopic distributions.

As a peak detection threshold is designed to efficiently separate the signal containing regions of the spectrum from those that contain noise, these thresholds can also be used for the purposes of spectral compression. In this mode of use, the noise regions in the spectrum are removed from the data with the aim of reducing the required storage space or the band-width required to transmit them. A high performing peak detection threshold will also provide a good threshold for spectral compression as it will remove regions of the spectrum containing no information whilst keeping those regions that do contain information.

METHODS

EXPERIMENTAL

Materials. Protein standards were purchased from Sigma Chemical Co. (St. Louis, MO). Methanol, water, and formic acid were purchased from Fischer Chemicals (Zurich, Switzerland) and were LC-MS or mass spectrometry grade. For native MS, samples were first desalted using Micro Bio-Spin 6 columns (Biorad)

Mass Spectrometry. Top-down electron capture dissociation (ECD) mass spectra were acquired with a Bruker Solarix 12 T FT-ICR mass spectrometer (Bruker Daltonics, Bremen, Germany). Myoglobin (2 μM in 50:50:0.1 methanol:water:formic acid) and alcohol dehydrogenase (5 μM in 100 mM ammonium acetate) were typically ionised by electrospray (4.0 kV, 200μL h^{-1}). For myoglobin, the 18+ charge state was isolated prior to ECD with accumulation time of 750 ms. For ECD, the cathode current was set to 1.5 A, the pulse length was 20ms, with a bias voltage of 1.5 V and the lens voltage of 15 V. For ECD analysis of native alcohol

dehydrogenase, ions were accumulated for 500 ms and subjected to ECD without prior isolation. ECD pulse length was 25 ms with 1.5 V bias and 15 V lens.

DATA ANALYSIS

The first step in the Autopiquer algorithm is to select a window width. We have found that for most spectra the window width should be 3 *m/z*. This selection of window width is one of the few required inputs in order for the Autopiquer algorithm to work. We envisage that, for most users, this value will never need to be changed and, compared to the other algorithms (THRASH, *n*-Sigma and RMS), the Autopiquer method is relatively insensitive to changes in this value.

A freely moving window (i.e. one that is centered upon every point across the spectrum in turn) would produce good results, but would be computationally expensive. Instead, the window used in Autopiquer moves by the pre-determined window width and the threshold derived for that window is used to apply to every point within the window. However, to ensure that peaks that cross window boundaries are properly dealt with, the spectral section that is used to calculate the autocorrelation for the window is the window width plus an additional half window width on either side (i.e. the total spectral section used to calculate the threshold for level for each window is two window widths wide). Therefore, the sections actually overlap, as shown in Figure 1.

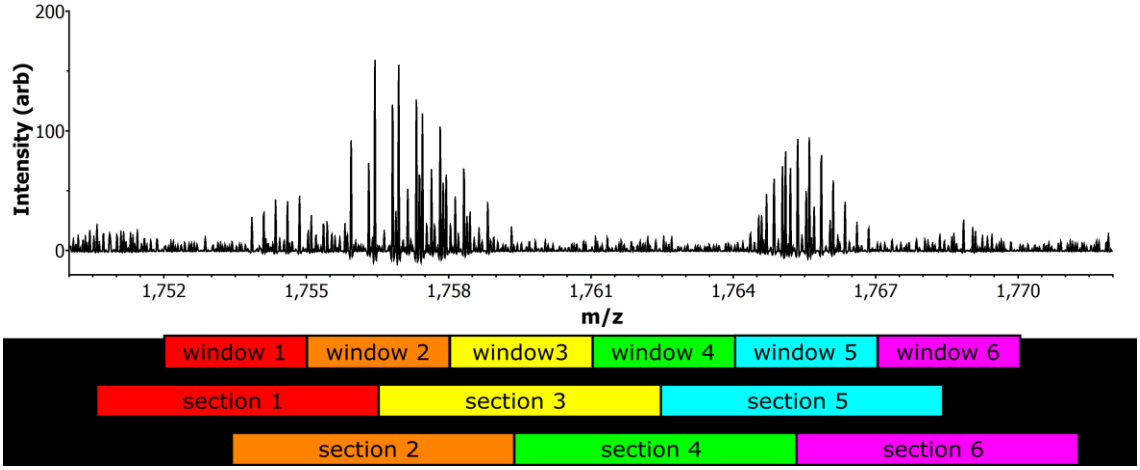

Figure 1 – Showing the mass ranges used for six adjacent windows across a portion of a mass spectrum generated by ECD fragmentation of native state alcohol dehydrogenase. The threshold value for each window would be calculated from the relevant spectral section (same number and colour as the window) – note that the sections overlap.

Figure 1 shows a small region the mass spectrum obtained during ECD top-down fragmentation of alcohol dehydrogenase (ADH) from *Saccharomyces cerevisiae*. ADH is composed of a homo-tetrameric assembly of subunits and has a molecular weight of 147 kDa. ADH has previously been investigated by native ECD-MS, as reported by Gross and coworkers[16] and Loo and coworkers.[17]

The spacing between points in mass spectra is usually not constant across the mass range. Therefore to make it easier to calculate the autocorrelation spectrum of each window the Autopiquer algorithm uses linear interpolation, within each spectral section, to generate a resampled spectral section where the points are equally spaced in the mass dimension and where the spacing equals the minimum mass spacing in the section (just that section, not the entire spectrum) or 1×10^{-6} times the lowest mass of the section, whichever is the greater. This coercion is used because for FT-MS techniques the number of points within a spectral section of fixed mass width at the very low mass end of the spectrum may be very large and calculating the threshold using the full autocorrelation of these regions would be overly time consuming. For

spectra exhibiting the very highest resolutions (for example spectra showing isotopic fine structure), this coercion limit can be reduced to 1×10^{-9} times the lowest mass of the section. However, this change would only be required for specialist users.

Within each one of the resampled spectral sections, the autocorrelation of the spectrum is calculated. The peaks in the autocorrelation spectrum correspond to the $\Delta\, m/z$ between peaks in the mass spectrum – i.e. approximately equal to the reciprocal of the charge state. Only that portion of the autocorrelation spectrum that could contain peaks resulting from reasonable isotopic peak spacings (or their harmonics) is considered in later steps. The lower limit of the region of interest of the autocorrelation spectrum is taken as a lag ($\Delta\, m/z$) of 0 and the upper limit is set to a lag of 2.25. The $\Delta\, m/z$ 2.25 upper limit of the autocorrelation region of interest is set because some isotope distributions from singly charged ions can produce a strong autocorrelation at $\Delta\, m/z = 2$. This limit is intended to ensure that this autocorrelation peak, if present, is included in the threshold estimation. When the autocorrelation spectrum is referred to below, this is intended to be limited to only this region of interest where $0 \leq \Delta\, m/z \leq 2.25$.

In the region of interest in the autocorrelation spectrum, the number of points where the autocorrelation value is ≤ 0 are counted, as a proportion of the total number of points in that region.

Next, a threshold, which is set at the minimum intensity value of the mass spectrum within the window, is applied to the resampled mass spectral section; any point in the resampled spectrum less than that threshold is set to zero intensity. The autocorrelation spectrum is calculated again

and the number of points (described above) in autocorrelation spectrum that are ≤0 are again counted. By iteration, we find the threshold level that, being applied to the resampled mass spectral section, results in a certain proportion of the points in the autocorrelation spectrum being ≤0. This proportion will vary across the spectrum as a consequence of both the spectral resolution and the number of spectral points per peak varying as a function of mass, and is calculated for every window. The width of the highest peak in the autocorrelation spectrum is measured by determining the number of points between the local minima on either side of the peak maximum and compared to the total length, in points, of the region of interest of the autocorrelation spectrum.

The target number of the points in the autocorrelation spectrum to be ≤0 is set as the same number as the width (in points, from local minimum to local minimum) of the highest peak in the autocorrelation spectrum for that section. This target was developed to reconcile two requirements: that the threshold level be set to be responsive to the varying resolving power and spectral point densities exhibited in each spectral region and from each mass spectrometer (with their different relationships between mass and resolution and spacings between spectral points), and that the measures used to control the adaptation of the threshold be robust and simple to calculate from the autocorrelation spectrum. Various approaches were tested and this method proved to reliably meet our requirements of performance, adaptability and ease of implementation.

The peak width (in terms of number of datapoints) of mass spectral peaks (in isotopic distributions) in the mass spectrum is closely related to the peak width (in points) in the corresponding autocorrelation spectrum. The two peak widths (in terms of the number of points) will not be exactly the same because the peak widths and spacing in the mass spectrum will only rarely be an

exact multiple of the point spacing. The peak width is not dependent on the charge state of the ion. For a high charge state ion, there may be many peaks in the autocorrelation spectrum and the Autopiquer algorithm is intended for use on spectra that are baseline resolved (or close to it). At the lowest intended resolution, the peak tails in the autocorrelation spectrum will just touch or slightly overlap. Therefore at this limit, if the autocorrelation spectrum (in the region of interest) comprised contiguous peaks, if more than a peak width of the autocorrelation spectrum was ≤0 then this could be because one (or more) of the autocorrelation peaks had been lost (normally the one at the highest $\Delta m/z$) – consequently one could infer that the threshold in the mass spectrum would have been set so high that some of the isotopic information in the spectrum has been lost and the threshold level should be reduced. Note – the autocorrelation spectrum does not generally have a flat baseline across the region of interest with this baseline tending to show a decrease towards higher $\Delta\, {}^{m}/_{z}$. This is why, if regions of the autocorrelation spectrum are ≤0, entire peaks can be lost rather than simply cutting through the valleys between peaks. This effect is illustrated in the Supplementary information section S 1.

As neither the noise nor the peak width are affected by the charge state, the level that is the correct noise threshold for high charge state peaks will also hold true for lower charge state peaks in the same region.

Additionally, as the peak width in mass spectra generally increases as a function of mass, the peak width within any window is not allowed to be lower than the peak width determined for lower mass windows. This check is to prevent issues in spectral regions where there are no peaks – the highest peak in the autocorrelation spectrum for these regions will be from random noise and is

most likely to be narrower than real peaks detected in lower mass regions of the spectrum. If the threshold was set using this peak width then the detection threshold would be set too low and many noise peaks would be detected in that region.

This method (and the others that were investigated as part of this research but which proved less successful) for detecting the noise level was developed using artificially generated mass spectral regions combining single (or overlapping) isotopic peak clusters were generated using the mercury algorithm[18,19] and spectral noise, generated using random number generation. In these artificial spectral regions, it was possible to have complete control over the effective spectral sampling rate, peak signal to noise ratios, resolution, ion charge state and the presence of overlapping peak distributions at different charge states. Once the method had been developed and successfully tested against the artificial spectral regions, it was applied to real spectra.

Having optimized the threshold level that results in the correct proportion of the autocorrelation spectrum being ≤ 0, this threshold level is then applied to the window at the core of the spectral section.

This process provides an intensity level, in each spectral window, below which the spectrum is apparently locally unstructured. If there are regularly spaced peaks in the mass spectrum then these will be reflected by peaks in the autocorrelation spectrum. We iterate to find a threshold level in the mass spectrum that preserves as much of this structuring as possible, whilst removing unstructured data. This approach can then be used as a peak detection threshold. It can also be used for data compression purposes – by deleting the portions of the spectrum below that threshold.

As the algorithm is estimating a level for the noise threshold, and the noise has some statistical distribution, there will be a proportion of noise peaks that still protrude above this level. To reduce the occurrence of these in the detected peak list, we have found that raising the applied peak detection threshold to a level that is some multiple of the detected noise threshold works well – i.e. setting a minimum signal to noise ratio (SNR). The signal mean (μ) is estimated from the mean of all points in the window that are less than the Autopiquer estimated noise threshold (l) for that window. The applied SNR threshold (T) is calculated as $T = x \times (l - \mu) + \mu$, where the SNR level ($x$) is commonly set to a value 1.5 to detect a useable peak list. Setting the value of the SNR to larger values (e.g. 3 or 5) can be useful in the event that one wishes to generate a high confidence peak list – for example for assigning peaks that will be used to generate an internal calibration function.

Once the threshold is calculated for every window, the threshold to use for every point in the mass spectrum can be easily derived. As a final step, and if desired, it is possible to smooth the steps in the threshold. A sigmoid spline can be applied to the thresholds in the last quarter of one window and the first quarter of the next; for all windows. This spline is based on the standard sigmoid function and is calculated, for each step by

$$F_p = \frac{T_{w2} - T_{w1}}{1 + e^{-p}} + T_{w1} \qquad \mathbf{1}$$

Where T_{w1} and T_{w2} are the thresholds for the windows before and after the step respectively, F_p is the calculated sigmoid spline and p is the point position in the mass spectrum counting the first point in the later window as $p = 0$ and scaled across the splined range such that $-6 \leq p \leq 6$.

We have programmed this algorithm as a sub-vi in National Instrument LabVIEW (Austin, Tx, USA) using the NI supplied autocorrelation function with unbiased normalization. However the algorithm could be easily programmed using any language.

A flow diagram summarizing the algorithm is shown in Figure 2.

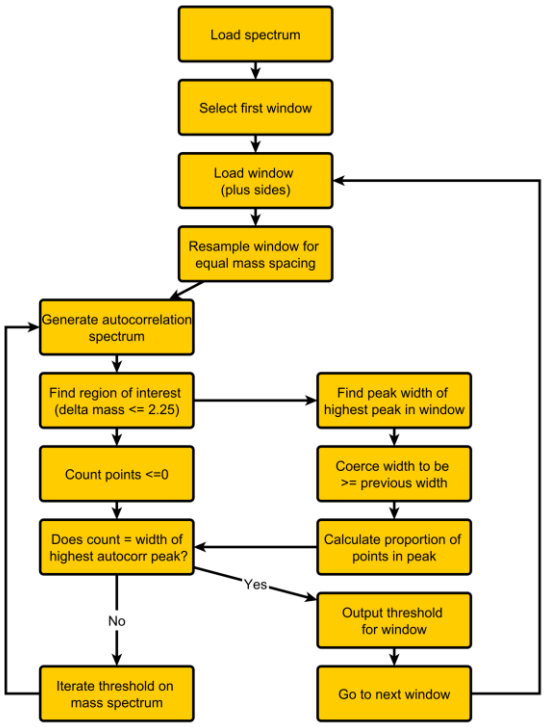

Figure 2 – Flow diagram summarizing the Autopiquer algorithm.

RESULTS & DISCUSSION

Figure 3 illustrates some of the process of optimizing a threshold for a window within a spectrum. Figure 3 (a) shows the spectral portion of the fully apodized FT-ICR mass spectrum of the electron capture dissociation of native state alcohol dehydrogenase from *Saccharomyces cerevisiae* (Sigma-Aldrich St Louis, MO, USA), containing the window bracketing the c'_{29}^{4+} fragment. We

have used ECD datasets to test the algorithm, because ECD is well known to be an inefficient fragmentation process which results in complex spectra characterized by signals of low S/N – ideal to test the utility of the algorithm. The spectrum was processed to absorption mode using the Autophaser method, applying a full apodization (F=0.5) to generate a baseline deviation free mass spectrum.[20-23] The superimposed isotopic distribution for the fragment was generated using the method described previously.[24] Figure 3 (b) shows a series of autocorrelations of the spectral section shown in (a) as the iteration towards the optimum threshold level progresses. The peaks in the autocorrelation spectrum are regularly spaced at $\Delta m/_z = 0.25$, indicating that the distribution is quadruply charged. As the threshold increases from iteration to iteration, the offset and gradient of the calculated autocorrelation spectrum is progressively removed, but the peaks in the autocorrelation spectrum remain, indicating that the information within the spectral section remains. The final optimized threshold is shown in Figure 3(a). The spectral mean is calculated for this portion by taking the mean of all points that are less than the optimized threshold level.

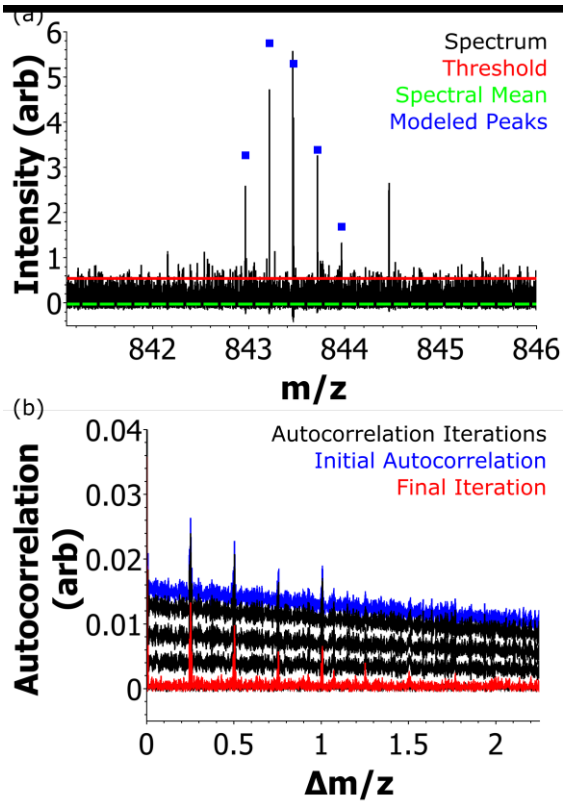

Figure 3 – (a) Spectral portion of mass spectrum of the electron capture dissociation of native state alcohol dehydrogenase, containing the window bracketing the c'29 4+ fragment, showing the Autopiquer optimized threshold (red line), the residual spectral mean (green dashed line) and the position of the modelled isotope distribution (blue points). (b) Showing the region of interest in the autocorrelation spectra of the mass spectral portion shown in (a). The initial iteration is shown in blue and the final iteration in red.

To illustrate the performance of the Autopiquer algorithm for setting a peak detection threshold across a spectrum we have compared it to the performance of the noise level estimation algorithm described by Horn et al., [6] (this algorithm was described alongside the famous THRASH algorithm and consequently, for conciseness, we have labelled it the THRASH threshold) and to the well-known root-mean squared (RMS) and n-Sigma methods (in this case, we have set $n = 2$). We have not included the peak detection methods used in commercial software from the instrument manufacturers as the algorithms by which these operate are not known and so it is

difficult to provide an objective test. However, in our experience, the difficulties that we find in using the published algorithms that we have used for illustration here are also found when using the peak detection algorithms available in commercial software.

For this test, all peak detection threshold estimate algorithms have been set to use the same window width (3 Da). The test spectrum (shown in Figure 4) is a top-down electron capture dissociation spectrum of denatured horse heart myoglobin (Sigma-Aldrich) collected on an FT-ICR MS and displayed in absorption mode. Equine myoglobin, in denatured *apo* form (having lost the heme group), is a ~17KDa protein and is now commonly used as a simple test compound for top-down ECD experiments. Previous work on the top-down characterization of horse heart myoglobin by ECD on FT-ICR MS was presented by Pan *et al* [25] and Mikhailov and Cooper.[26]

Figure 4 (a), shows the performance of the four threshold algorithms in a low mass region of the spectrum. Peak resolution here is high, relative to the peak density, meaning that the peaks are narrow and well-spaced. Under these conditions the Autopiquer (with $SNR = 1.5$), THRASH and *n*-Sigma methods all return a reasonable estimate of the peak detection threshold. The RMS method returns a threshold that, by eye is apparently too low.

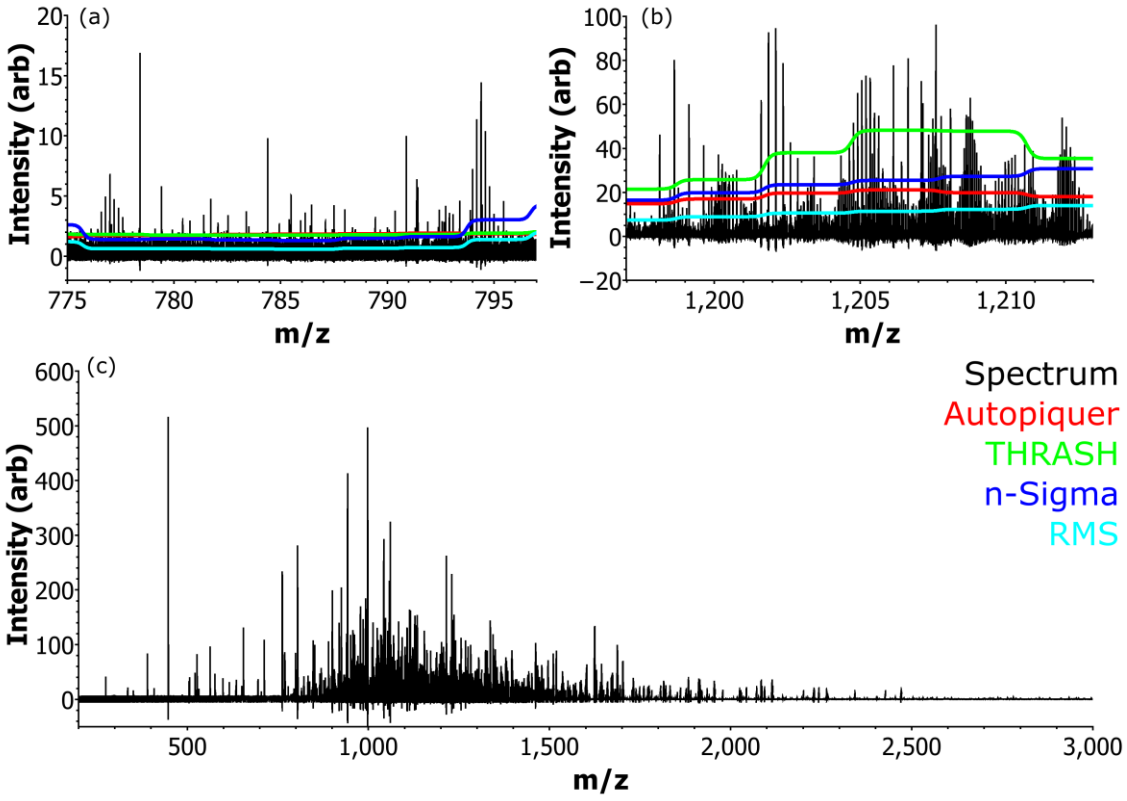

Figure 4 – Showing peak detection thresholds being set for an absorption mode FT-ICR MS spectrum of the electron capture dissociation product ions of denature horse heart myoglobin. The complete spectrum is shown in panel (c), bottom row. Panels (a) and (b), top row, show the peak detection thresholds generated by different algorithms in two regions of interest – one with low peak density and with high peak density. For all methods, the window width was 3 Da. The spectrum is shown in black and the thresholds calculated by the Autopiquer, THRASH, *n*-Sigma and RMS methods are shown in red, green, blue and cyan respectively.

In more peak dense regions of the spectrum, in this case at higher mass, as illustrated in Figure 4 (b), the THRASH (green line) and *n*-Sigma (blue line) algorithms return peak detection thresholds that are too high and miss many peaks. The Autopiquer (red line) and RMS (cyan line) methods return thresholds that are more reasonable.

Only the Autopiquer algorithm, of the four under test, provides an adequate peak detection threshold across both peak sparse and peak dense portions of the spectrum. The other three

methods did provide a useful estimation of the threshold in one region, but not in the other. This example highlights why these algorithms may prove difficult to use – it is often challenging, even with a single spectrum, to define a single setting that will robustly detect peaks across the complete spectrum. An additional discussion of the effects of peak density on the thresholds set by the different techniques, and using synthetic mass spectral regions (to allow specific control of the peak density), is provided in the Supplemental Information S 5.

When the peak lists generated using the four different algorithms are assigned, one would expect that the peaks list resulting from the RMS method would result in a high number of false assignments in the low mass region whilst the peak lists resulting from the THRASH and *n*-Sigma methods would suffer from many missed assignments in the higher mass region. To test this hypothesis, we developed a simple peak assignment tool that generates a library of potential fragment ion masses from a given protein sequence (including post-translational modifications where required), calculating the position and relative intensity (normalized to the base peak in each isotopic cluster) of all isotopologue peaks for each fragment using previously described methods.[18,19,27] A series of peaks is assigned to a particular fragment only if all isotopologue peaks that the library indicates should be detectable, given the signal-to-noise ratio of the base peak in the distribution and the noise level calculated by the peak detection threshold algorithm, are found in the spectrum within the user defined mass error limits.

Using this approach, we assigned the peak lists generated by the four different thresholding methods for the horse heart myoglobin spectrum described above, with the results shown in Figure 5. The mass error limit for assignment was set to ± 6 ppm and the fragments classes were restricted to *a*, *b*, *c*, *y* and *z* type protein fragment ions. As expected, the THRASH and *n*-Sigma thresholds result in poor assignment rates in peak dense portions of the spectrum and the RMS method results

in a very large number of false assignments in the low mass region. The *n*-Sigma method performs the best of the common techniques (in this example – based on both the overall sequence coverage and also on the number of assigned fragments supporting that sequence), but its assignment rate (75%) is lower than for the Autopiquer detected peak list (assignment rate – 88%) implying that there are many peaks that are still missed by the use of the *n*-Sigma approach that are detected by the Autopiquer method under similar conditions. This can be seen by the number and density of points in the Autopiquer panel (of Figure 5) compared to the results of the other methods.

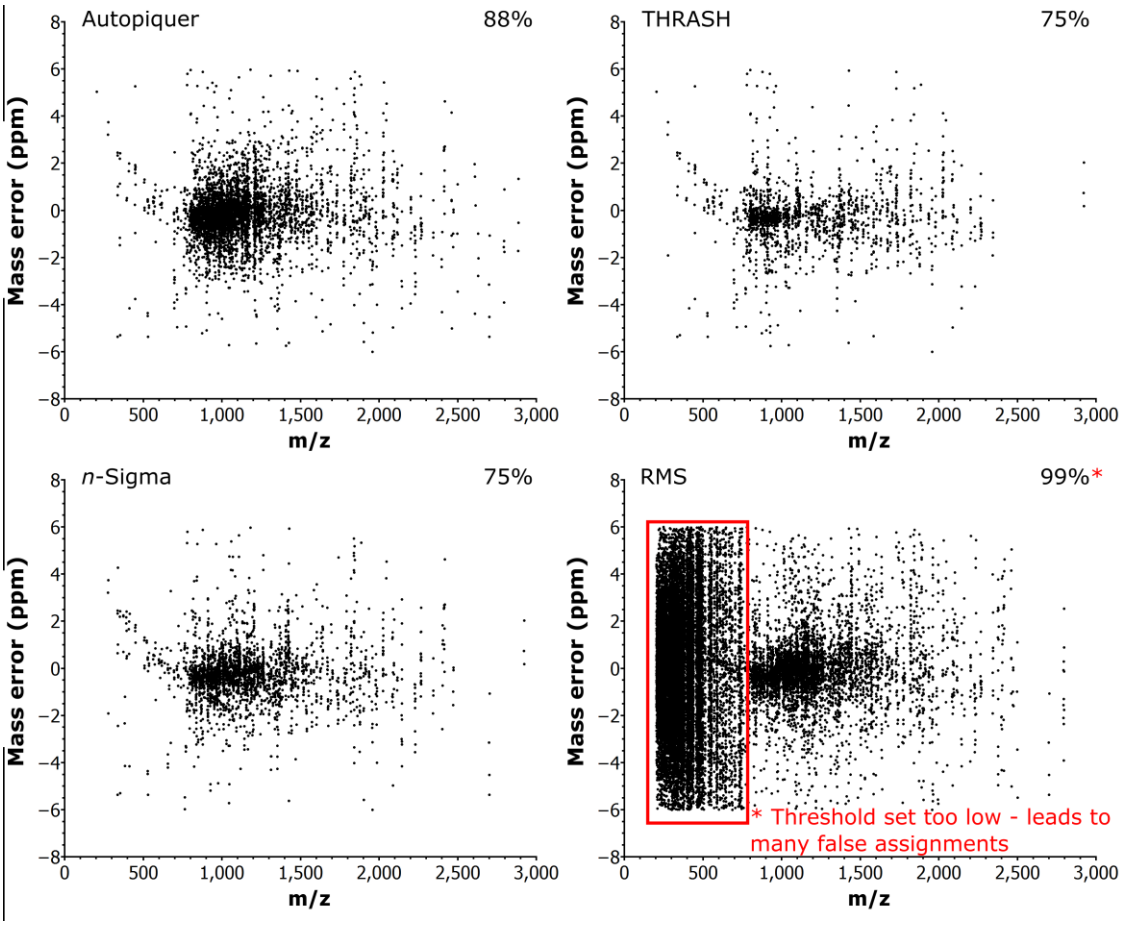

Figure 5 – Assignments (within ±6 ppm) for the peak lists generated by using the four different peak detection thresholding methods with all using the peak assignment method described in the text.

In order to further compare the performances of the different techniques, we use a measure of the "peak efficiency" calculated as the ratio of the percent protein sequence coverage (ignoring false positives for the sake of simplicity) to the number of detected peaks (in thousands) in the peak list. For the myoglobin spectrum, the Autopiquer and THRASH methods clearly outperform the *n*-Sigma and RMS methods, as shown in Table 1. The higher the peak efficiency value, the smaller the proportion of unassigned peaks and hence the smaller the potential for missed assignments. Compare, for example the fact that the THRASH and *n*-Sigma approaches result in the same proportion of sequence coverage, but that coverage was derived from more than 44,000 peaks from the *n*-Sigma method but from less than 9,000 peaks from the THRASH method. This is reflected in the peak efficiency of the THRASH method being almost five times higher than for the *n*-Sigma method. The peak list generated by the Autopiquer algorithm contains the fewest peaks but provides the highest sequence coverage in this example; consequently, the Autopiquer algorithm presents the highest peak efficiency.

Method	%	No of peaks	No of Assigned peaks	Peak efficiency	% of Spectrum Above Threshold
Autopiquer	88	8960	2890	9.8	1.34
THRASH	75	8970	2007	8.3	1.06
n-Sigma	75	44039	2185	1.7	3.19
RMS	99	152200	18167	0.65	15.97

Table 1 – Metrics of peak detection and spectral compression by four different peak threshold methods.

Besides the absorption mode FT-ICR MS spectra used as examples here, the performance of the Autopiquer algorithm has also successfully tested against magnitude mode spectra (the traditional, lower performance output mode for FT-MS data) from FT-ICR MS and on spectra from lower resolution mass spectrometers (e.g. MALDI TOF and LC-QTOF MS). For example, *Figure 6* illustrates the difference in performance of the Autopiquer algorithm and the *n*-Sigma approach

for a LC-QTOF mass spectrum. This data was collected on a Bruker maXis Impact Q-TOF instrument and shows peaks related to components in a microbiological growth medium, averaged from a portion of the complete chromatographic time. As in the other examples, the Autopiquer algorithm successfully identifies the noise level (using the standard setup). The *n*-Sigma (where n = 2) threshold is shown for comparison (for a window width of 3 Da). Autopiquer has also been shown to be a successful approach for spectra that have been baseline corrected and smoothed using continuous wavelet transform methods. A MALDI-TOF example of this is provided in Section S 2 in the Supplementary Information.

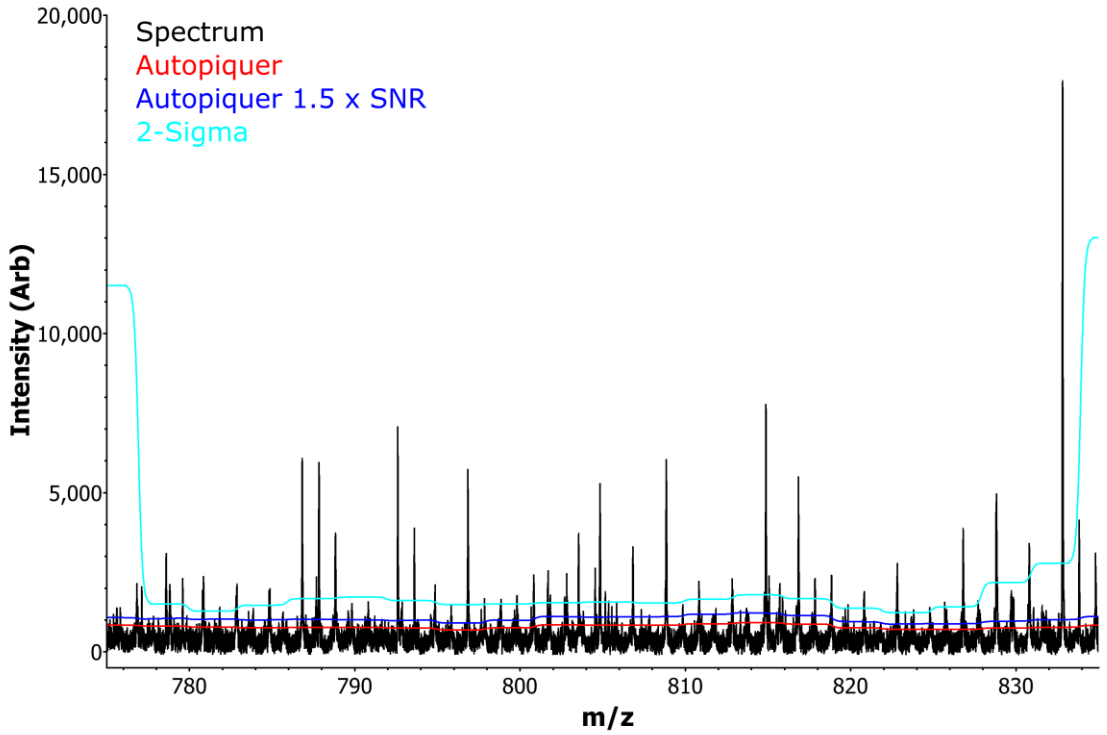

Figure 6 - Mass spectrum created by averaging part of a liquid chromatography mass spectrometry data-file, recorded on a Bruker maXis Impact Q-TOF, where the sample was a microbiological growth medium. Three different thresholds are shown for comparison: the Autopiquer threshold, the 1.5 × SNR Autopiquer threshold and the n-Sigma threshold (where n = 2).

Spectral compression

Any of the four threshold methods described above could also be used for spectral compression purposes. However, regions of the spectrum in the examples above, where peaks were missed, because the thresholds from some of the algorithms were too high, would now be permanently removed from the data; this could then not be recovered and would result in permanent information loss, and so is an outcome to avoid. Regions, in the examples above, where the peak detection threshold was set too low would now cause unwanted noise regions to be incorporated into the compressed spectrum, increasing the file size. However, as any signal in those regions could be retrospectively extracted, the information in the spectrum is still present, so this outcome is less deleterious than the removal of actual signal, albeit at the cost of increased data volume.

Any metric for the success of spectral compression must present the inevitable compromise between the extents of data reduction (good) versus information reduction (bad). The extent of data reduction available in a spectrum will depend on the peak density and on the spectral resolution and so will vary strongly between spectra and different instrument classes. The extent of information reduction (as a consequence of spectral compression) is difficult to measure accurately in real spectra as you do not know the true number of peaks or their identities. However, we can illustrate the performance of Autopiquer for spectral compression on a single spectrum, using the same metrics as were used for peak detection, and shown in Table 1.

The extent of data reduction can be estimated either by the number of detected peaks (for the situation where the intent was to record only peak centroids in a reduced spectrum) or by the percentage of the original spectral data points that lie above their respective thresholds (where the

intent was to record the spectral profile above threshold). The proportion of spectral information remaining after compression cannot be determined accurately but the protein sequence coverage and number of assigned peaks can be used to estimate the relative success of the different thresholding techniques in this example.

Based on these measures, the Autopiquer algorithm provides the highest performing spectral compression threshold for the example spectra (both the myoglobin example here and in the cytochrome c spectrum in Section S 3 in the Supplementary Information) because it both retains the smallest proportion of the original spectral data and the highest proportion of the original information. Using the THRASH algorithm to generate a threshold for spectral compression would result in a similar file size to the compressed data using the Autopiquer threshold, but the THRASH algorithm also results in the loss of more of the actual information. The *n*-Sigma method, in the example above, retained approximately the same proportion of the information as the THRASH method but would result in a data volume approximately three to five times larger. It is hard to accurately estimate how much of the spectral information is maintained above the threshold from the RMS example. However, the compressed data volume would be considerably higher than for the other methods.

The Autopiquer algorithm has an additional advantage. Unlike the other thresholding methods described in this paper, the Autopiquer algorithm can be successfully applied to previously compressed data as it generates a threshold based on the spectral structure and does not need noise to be present from which to gather statistical values to use to calculate the threshold. Therefore, not only can it used to detect peaks in data that has been previously thresholded by another means, but it can also be used to check the level at which that threshold was applied. If that threshold was applied at a much higher level than the threshold that the Autopiquer algorithm would have

returned, then the proportion of points in the autocorrelation spectrum of each region of interest that are $\leq O$ will be higher than expected. In order to perform this test, it is necessary to estimate the total peak width for the data based on fitting an expected peak shape to the residual peak tops that will be present in the thresholded data. However, by this means, the Autopiquer algorithm can be used to estimate if the thresholding methods used by other software, where the method by which the threshold was set may not be clear, could be resulting in excessive information loss from the saved data.

Window length

In the examples above, all threshold algorithms have been set to use the same window length, in order to reduce the number of variables in each example. The window length does have an effect on the thresholds set by all the algorithms. Even following optimization of the windows lengths of each algorithm, the Autopiquer algorithm still provides the best performance. We provide an example of this, using the THRASH algorithm, in Section S 4 in the Supplementary Information.

CONCLUSIONS

Our aim of providing an improved method of generating a peak detection/spectral compression threshold has resulted in a new algorithm we call Autopiquer. Autopiquer optimizes a threshold level that removes as much noise as possible from across a mass spectrum whilst maintaining as much of the isotopic peak structures as possible.

Autopiquer has been tested against well-established threshold estimation algorithms and shows improved performance by every measure when we have tested it. Not only does the method apparently produce the most complete peak lists (in terms of attempting to capture the real

information in a spectrum) but it also minimizes the number of noise peaks included – thereby reducing the proportion of false positive hits. Furthermore, the method only requires the user to adjust one variable, the SNR ratio. And, even then, in almost all spectra we have processed by this method, the value of this never needs to be changed from the default setting of 1.5. Therefore this algorithm for producing a peak detection threshold requires much less user interaction and skill in order to process mass spectra and is suitable for high throughout processing of data – for either peak detection or spectral compression.

In practice, we have found this peak detection method routinely outperforms commercially available peak detection methods in that it takes considerably less user skill (and time) to produce a peaks list that is at least as good as, and usually much better than, the peak lists returned by the commercial programs in application areas including proteomics, enzymology and lipidomics. We note that the Autopiquer algorithm is only intended to work on high-resolution mass spectra, where isotopic structure in the mass spectrum is available, because the isotopic spacing of peaks is used to help optimize the threshold. Finally, it should also be noted that the Autopiquer method should, in principle, not be limited to analysis of mass spectra, but could in fact be extendable to any form of spectroscopy that produces regularly spaced peaks in the spectra.

Supporting Information Available: Examples of the use of Autopiquer on mass spectra from different instrument types and sample classes, top down sequence coverage against other proteins, and the effect of window length on threshold level using other algorithms.

ACKNOWLEDGEMENTS

DJC would like to thank the University of Edinburgh for the award of Chancellor's Fellowship. SH would like to thank the EaStCHEM School of Chemistry for funding.

REFERENCES

1. Du, P., Kibbe, W. A., Lin, S. M.: Improved Peak Detection in Mass Spectrum by Incorporating Continuous Wavelet Transform-Based Pattern Matching. Bioinformatics **22**, 2059-2065 (2006).

2. Mantini, D., Petrucci, F., Pieragostino, D., Del Boccio, P., Di Nicola, M., Di Ilio, C., Federici, G., Sacchetta, P., Comani, S., Urbani, A.: LIMPIC: A Computational Method for the Separation of Protein MALDI-TOF-MS Signals from Noise. BMC Bioinformatics **8**, 1-17 (2007).

3. Meuleman, W., Engwegen, J. Y. M. N., Gast, M. W., Wessels, L. F. A., Reinders, M. J. T.: Analysis of Mass Spectrometry Data using Sub-Spectra. BMC Bioinformatics **10**, 1-9 (2009).

4. March, R. E.; Todd, J. F. Practical Aspects of Trapped Ion Mass Spectrometry, Volume IV: Theory and Instrumentation; CRC Press, Boca Raton, FL, USA, 2010.

5. Zhurov, K. O., Kozhinov, A. N., Fornelli, L., Tsybin, Y. O.: Distinguishing Analyte from Noise Components in Mass Spectra of Complex Samples: Where to Cut the Noise? Anal. Chem. **86**, 3308-3316 (2014).

6. Horn, D. M., Zubarev, R. A., McLafferty, F. W.: Automated Reduction and Interpretation of High Resolution Electrospray Mass Spectra of Large Molecules. J. Am. Soc. Mass Spectrom. **11**, 320-332 (2000).

7. R. Chellappa, P. Sinha, P. J. Phillips.: Face Recognition by Computers and Humans. Computer **43**, 46-55 (2010).

8. Hussong, R., Tholey, A., Hildebrandt, A.: Efficient Analysis of Mass Spectrometry Data using the Isotope Wavelet. AIP Conference Proceedings **940**, 139-149 (2007).

9. Hussong, R., Gregorius, B., Tholey, A., Hildebrandt, A.: Highly Accelerated Feature Detection in Proteomics Data Sets using Modern Graphics Processing Units. Bioinformatics **25**, 1937-1943 (2009).

10. Köster, C.: Mass Spectrometry Method for Accurate Mass Determination of Unknown Ions. Mass Spectrometry Method for Accurate Mass Determination of Unknown Ions. (2001).

11. Senko, M. W., Beu, S. C., McLafferty, F. W.: Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. J. Am. Soc. Mass Spectrom. **6**, 229-233 (1995).

12. Lange, E., Gropl, C., Reinert, K., Kohlbacher, O., Hildebrandt, A.: High-Accuracy Peak Picking of Proteomics Data using Wavelet Techniques. Pacific Symposium on Biocomputing **11**, 243-254 (2006).

13. Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., Kobayashi, R.: Feature Extraction and Quantification for Mass Spectrometry in Biomedical Applications using the Mean Spectrum. Bioinformatics **21**, 1764-1775 (2005).

14. Palmblad, M., Westlind-Danielsson, A., Bergquist, J.: Oxidation of Methionine 35 Attenuates Formation of Amyloid Beta -Peptide 1-40 Oligomers. J. Biol. Chem. **277**, 19506-19510 (2002).

15. Catherman, A. D., Skinner, O. S., Kelleher, N. L.: Top Down Proteomics: Facts and Perspectives. Biochem. Biophys. Res. Commun. **445**, 683-693 (2014).

16. Zhang, H., Cui, W., Wen, J., Blankenship, R. E., Gross, M. L.: Native Electrospray and Electron-Capture Dissociation in FTICR Mass Spectrometry Provide Top-Down Sequencing of a Protein Component in an Intact Protein Assembly. J. Am. Soc. Mass Spectrom. **21**, 1966-1968 (2010).

17. Li, H., Wongkongkathep, P., Van Orden, S. L., Ogorzalek Loo, R. R., Loo, J. A.: Revealing Ligand Binding Sites and Quantifying Subunit Variants of Noncovalent Protein Complexes in a Single Native Top-Down FTICR MS Experiment. J. Am. Soc. Mass Spectrom. **25**, 2060-2068 (2014).

18. Rockwood, A. L., Van Orden, S. L., Smith, R. D.: Rapid Calculation of Isotope Distributions. Anal. Chem. **67**, 2699-2704 (1995).

19. Rockwood, A. L., Van Orden, S. L.: Ultrahigh-Speed Calculation of Isotope Distributions. Anal. Chem. **68**, 2027-2030 (1996).

20. Kilgour, D. P. A., Wills, R., Qi, Y., O'Connor, P. B.: Autophaser: An Algorithm for Automated Generation of Absorption Mode Spectra for FT-ICR MS. Anal. Chem. **85**, 3903-3911 (2013).

21. Kilgour, D., Neal, M. J., Soulby, A. J., O'Connor, P. B.: Improved Optimization of the Fourier Transform Ion Cyclotron Resonance Mass Spectrometry Phase Correction Function using a Genetic Algorithm. Rapid Commun. Mass Spectrom. **27**, 1977-1982 (2013).

22. Qi, Y., Li, H., Wills, R. H., Perez-Hurtado, P., Yu, X., Kilgour, D. P. A., Barrow, M. P., Lin, C., O'Connor, P. B.: Absorption-Mode Fourier Transform Mass Spectrometry: The Effects of Apodization and Phasing on Modified Protein Spectra. J. Am. Soc. Mass Spectrom. **24**, 828-834 (2013).

23. Kilgour, D., Van Orden, S. L.: Absorption Mode Fourier Transform Mass Spectrometry with no Baseline Correction using a Novel Asymmetric Apodization Function. Rapid Commun. Mass Spectrom. **29**, 1009-1018 (2015).

24. Kilgour, D. P. A., Van Orden, S. L., Bao Quoc Tran, Goo, Y. A., Goodlett, D. R.: Producing Isotopic Distribution Models for Fully Apodized Absorption Mode FT-MS. Anal. Chem. **87**, 5797-5801 (2015).

25. Pan, J., Han, J., Borchers, C. H., Konermann, L.: Hydrogen/Deuterium Exchange Mass Spectrometry with Top-Down Electron Capture Dissociation for Characterizing Structural Transitions of a 17 kDa Protein. J. Am. Chem. Soc. **131**, 12801-12808 (2009).

26. Mikhailov, V. A., Cooper, H. J.: Activated Ion Electron Capture Dissociation (AI ECD) of Proteins: Synchronization of Infrared and Electron Irradiation with Ion Magnetron Motion. J. Am. Soc. Mass Spectrom. **20**, 763-771 (2009).

27. Kilgour, D. P. A., Van Orden, S. L., Tran, B. Q., Goo, Y. A., Goodlett, D. R.: Producing Isotopic Distribution Models for Fully Apodized Absorption Mode FT-MS. Anal. Chem. **87**, 5797-5801 (2015).